The use of Web Ontology Language (OWL) to Combine Extant Controlled Vocabularies in Biodiversity Informatics Appears Redundant

Roger Hyam (roger@hyam.net) July 2009

PESI WP4 Project Officer, Department of Botany, Natural History Museum, London.

Abstract

Implementation of PESI requires data to be combined from multiple source databases. Some of the shared fields in the source databases used different controlled vocabularies of terms. OWL DL was investigated as a mechanism to build an extensible, shared ontology of species occurrence terms that permitted the source database to continue using and extending their own vocabularies whilst formally mapping to a more generic shared vocabulary. The merits of this approach were explored and it was concluded that the building of such a complex mapping ontology probably wasn't worthwhile. The level of semantic complexity involved outweighed the costs of simply imposing a flat list of well defined terms onto data suppliers. The main problem with exiting vocabularies appear to be the overloading of terms. A candidate list of terms was proposed.

Introduction

Descriptive biology frequently states the nature of the occurrence of a species within a specified geographic region. At its simplest level this takes the form of a presents/absents/unknown statement. For example the flora of a country may list the provinces each species occurs in. Many publications go further and use other phrases to associate taxa with regions such as “Naturalised” or “Extinct”. This is an effective way of presenting summary information derived from numerous individual observations.

The Pan European Species dictionaries Infrastructure (PESI)[1] is an European Union funded project to unite the authoritative species name registers and nomenclators (name databases) and associated expertise networks that underpin the management of biodiversity in Europe. In its first instance PESI involves combining three key source databases Euro+Med PlantBase (EM)[2], European Register of Marine Species (ERMS)[3] and Flauna Europaea (FE)[4] to build a single, database driven portal to the taxonomy of European species. In the future there will be a requirement to incorporate further species databases from different regional focal points and taxonomic expert groups.

The three initial databases use different controlled vocabularies (lists of terms) for their respective occurrence status fields. The combined database, that drives the portal, therefore needs to have a unified list that combines these terms in a logically coherent way so that, for example, when a user searches for “Present” they find those records scored as “Naturalised” in one database and “Present” in another because “Naturalised” is seen as a sub category of “Present” in the combined list of terms. All three databases need to maintain their original lists of terms and possibly expand them because they are domain specific. Furthermore additional databases will be added in the future that may contain yet more terms.

1 http://www.eu-nomen.eu/pesi/
2 http://www.emplantbase.org/home.html
3 http://www.marbef.org/data/erms.php
4 http://www.faunaeur.org/

To solve this problem in a logically rigorous manner it was decided to use the Description Logics dialect of the Web Ontology Language (OWL) to build an Extensible Species Occurrence Ontology (ESOO). This approach was chosen for two key reasons. Firstly the primacy of the web in integration of biodiversity informatics data suggested the use of a technology tightly bound to it. Secondly, although the problem appears simple in the initial list of terms, the complexity was expected to grow rapidly and it was felt the use of inference would become important. It was also anticipated that the ontology itself may be useful in carrying out inference in combinations with other ontologies. This perceived need for inference ruled out the use of the Simple Knowledge Organisation System (SKOS)[5] which is expressed in the OWL Full dialect and is therefore not guaranteed to be computable.

Building ontologies of terms is a common problem not only within the biodiversity informatics field but more generally. It is hoped that a practical implementation of the approach will be informative when considering combining data for other controlled vocabularies notably habitat classifications, functional types and even geographic regions.

Terms Used In Source Databases

It is important to fully understand the terms used in the source databases prior to designing the ontology. Furthermore the act of designing the ontology can help elucidate an understanding of the subject domain.

The **ERMS database** does not record absence data but only tracks records of species occurrence in a region, an indication of whether it considers these records to be valid records and how certain it is about that validity. ERMS therefore recognises four possible states of presence. (Table 1)

Table 1: ERMS Occurrence States

Validity	Certainty	Interpretation
Valid	Certain	This record should be considered as good evidence of presence of the taxon in the region.
Valid	Uncertain	There is evidence of presence but some doubt about over the veracity of the information.
Invalid	Certain	This record asserts that the taxon is present in the region but the reviewing expert is certain it is wrong and it should not be used as evidence of presence. (This is not evidence of absence only a negation of a single record)
Invalid	Uncertain	This record asserts that the taxon is present in the region. The reviewing expert considers there is sufficient evidence to show that the record is probably wrong and should not be used as evidence of presence.

The **Fauna Europaea** database records presence and absence summaries for an area. There are four possible states for the occurrence field for a region (Table 2). Contributors have expressed a desire for more detailed codes to cover such things as migrants.

Table 2: Fauna Europaea Occurrence States

Term	Code	Interpretation

5 http://otto.w3.org/TR/skos-reference/

Present	P	There is at least one well documented record of the taxons presence in the area since 1600.
Doubtful	P?	The taxon is scored as being present in the area but there is some doubt over the evidence. The doubt may be of different kinds including taxonomic or geographic imprecision in the records.
Absent	A	The expert does not know of the existence of any records that assert the presence of the at taxon in this area.
<empty>	<empty>	The null condition. This record has not been scored.

The **Euro+Med PlantBase** database has a more complex schema that was originally based on the TDWG **Plant Occurrence and Status Scheme** (POSS)[6] standard. This standard consists of defining seven fields with a series of possible single letter values for each:

Field 1: Occurrence – Present (P), Assumed Present (S), Doubtfully Present (D), Extinct (E), Recorded as present in error (F).

Field 2: Native Status – Native (N), Assumed to be Native (S), Doubtfully Native (D), Formerly Native now extinct (E), Not Native (A), Recorded as Native in Error (F), No information (-), None of the above (U), Not Applicable (X).

Field 3: Introduction Status – Introduced (I), Assumed to be introduced (S), Doubtfully introduced (D), Formerly introduced now extinct (E), Not introduced (A), Recorded as introduced in error (F), No information (-), None of the above (U), Not applicable (X).

Field 4: Introduction Agency – Introduced by humans (M), Introduced by natural means (N), No Information (-), Not applicable (X).

Field 5: Cultivated Status – Cultivated outdoors (C), Cultivated indoors (I), Assumed to be cultivated (S), Doubtfully cultivated (D), Formerly cultivated now extinct (E), Not cultivated (A), Recorded as cultivated in error (F), No information (-), None of the above (U), Not applicable (X).

Field 6: Area Distribution Completeness – Distribution complete (C), Distribution incomplete (I), Not known whether distribution complete (U), Not applicable (X).

Field 7: World Distribution Completeness – Distribution complete (C), Distribution incomplete (I), Not known whether distribution complete (U).

E+M uses four fields each with a series of codes (Tables 3,4,5 & 6). Some of these codes are present for historical reasons, will not be used in future and will eventually be replaced but need to be accounted for – they have been deprecated. Many are little used. In addition to this E+M exports the values to a single (Table 7) that combines the fields omitting endemism.

Table 3: E+M Native Status Field Values

Code	Interpretation	Usage
A	Present: alien (definitely not native)	<1%
D	Present: doubtfully native (perhaps introduced only)	1%
E	Formerly native but presumably extinct	<1%
F	Absent but reported in error	2%

6 http://www.tdwg.org/standards/106/

N	Present: native	82%
Q	Presence questionable	1%

Table 4: E+M Introduction Status Field Values

Code	Interpretation	Usage
A	Definitely not introduced (deprecated)	< 1%
D	Present: doubtfully introduced (perhaps cultivated only)	< 1%
E	Formerly introduced but presumably extinct	< 1%
F	Absent but reported in error	< 1%
I	Introduced. If feasible, more precise categories are used such as	9%
I(A)	Adventitious (casual)	1%
I(N)	Naturalised	2%
I(P)	Problematic (degree of naturalisation uncertain)	1%
Q	Presence questionable	< 1%

Table 5: E+M Outdoor Cultivated Field Values

Code	Interpretation	Usage
A	Definitely not cultivated (deprecated)	< 1%
C	Present: cultivated	3%
D	Present: doubtfully cultivated (deprecated)	< 1%
E	Formerly cultivated but presumably extinct (deprecated)	< 1%
F	Absent but reported in error	< 1%
Q	Presence questionable (deprecated)	< 1%

Table 6: E+M Endemism Field Values

Code	Interpretation	Usage
C	distribution in Euro+Med area complete (endemic)	81%
I	distribution in Euro+Med area incomplete (not endemic)	18%
U	unknown whether distribution complete or not (deprecated)	1%

Table 7: E+M Actual Values Exported

Code	Interpretation	Usage
cultivated: C	Present: cultivated	3%
cultivated: F	Absent but reported in error	0%

introduced: D	Present: doubtfully introduced (perhaps cultivated only)	0%
introduced: E	Formerly introduced but presumably extinct	0%
introduced: F	Absent but reported as introduced present in error	0%
introduced: I	Introduced. If feasible, more precise categories are used such as	6%
introduced: I(A)	Adventitious (casual)	2%
introduced: I(N)	Naturalised	3%
introduced: I(P)	Problematic (degree of naturalisation uncertain)	0%
introduced: Q	Presence questionable	0%
native: D	Present: doubtfully native (perhaps introduced only)	1%
native: E	Formerly native but presumably extinct	0%
native: F	Absent but reported in error	2%
native: N	Present: native	82%
native: Q	Presence questionable	1%

Purpose of Occurrence Status Codes

Considering just these three databases and the POSS data standard reveals a considerable amount of semantic complexity. Manually mapping all the possible relationships between the codes and fields used is intimidating. There are numerous ways they could be mapped into a single formal framework. In order to make decisions as to which mapping may be optimal there needs to be a clear statement as to the purpose of occurrence status codes. A proposed solution can then be judged by its ability to meet these requirements effectively.

Occurrence statuses are fundamentally about the presence or absence of a taxon in a region.

Presence is usually based on evidence from observations made within the region but this begs a whole series of question: What evidence is needed to confirm presence? Does the observation have to be vouchered? If it is vouchered how certain do we need to be of its identity? Which taxonomic classification do we use? How long ago can observations have been made for them still to count as extant presence records? What about fossils? How long does the organism have to be present in the region for it to count as present? It is also possible to derive presence from other data. Some organisms are ubiquitous in certain environments. Obligate parasites or pollinators might indicate the presence of their associated taxa. Footprints and other marks may be taken as evidence of presence.

The scoring of **Absence** is often considered more problematic than scoring of presence. Some studies and databases do not record it at all. EMRS, for example, only tracks evidence for presence. It can be argue that it is not possible to prove a negative (a logical fallacy in which it is claimed that a premise is true only because it has not been proven false). There are, however, a number of situations where it is reasonable to assume absence of a taxon from a region – certainly with a similar level of certainty to many measures of presence. If a region is inhospitable to a species it may be assumed that the species does not exist in that region. An area composed entirely of sea water will lack freshwater species and certainly lack amphibians. It is also possible to combine evidence. An alpine calcicole plant can be assumed not to occur in a lowland acid bog especially if none of the plants usually associated with it occur there. This is certainly as good evidence for absence as the footprints in snow are evidence of presence of a particular large mammal.

Perhaps the largest source of absence records comes from controlled monitoring, particularly of birds, where a well defined set of observations are made according to a specific protocol and if the species is not observed it is taken to be absent.

There appears to be a great complexity of meaning even in the terms **present** and **absent**. Combining data from two sources may not be safe even when they have used the same terms as they may be based on very different evidence bases. Status codes are ultimately synoptic – they are summaries of underlying data and expert opinion based on the guidelines and context of a particular publication. As such they qualitatively different and can't reliably be combined between publications without loss of semantics. This is true even when we are considering only the terms present and absent let alone any of the more complex terms present in the POSS vocabulary and that are widely used for organisms like birds – such as Winter Feeding Migrant.

If there is a requirement to combine status terms between publications, as here, then this must be done through a level of abstraction or generalisation not by mapping directly from term to term. This is the approach that will be taken here.

Size of Vocabularies

Because POSS data standard has seven fields with a controlled list for each field the total number of combinations of field values is 233,280. A large percentage of these would be contradictory but no formal mechanism is in place to validate individual combinations. Leaving out the last two fields that deal with completeness of data – something that should perhaps be relegated to metadata – there are still nearly twenty thousand possible combinations.

As an indication of the significance of these sizes Voice of America can read the news in Special English, with a core vocabulary of only 1,500 words. This is not an entirely fair comparison as POSS could be thought of as a grammar consisting of separate words rather than individual words composed of seven letters but still the level of expressiveness is very high.

Inherited from POSS the E+M vocabulary uses multiple fields and therefore has the possibility of producing a very large number of combinations of values. This is currently 1,470 but drops to 288 if the terms marked for deprecation are removed. In actual fact these are mapped to just fifteen terms in a single field for export and only eight of these are actually used in the current data set.

Some audiences will need high levels expressiveness but for many applications the complexity presented by the POSS is a hindrance. Any solution has to be able to map from the complex to the simple.

OWL Modelling

The term ontology is somewhat overloaded. Here ontology is used synonymously with OWL file. A single OWL file contains a set of assertions concerning resources and their relationships. These include the declarations of classes, subclass relationships and equivalent classes. OWL files may also import other OWL files in which case the assertions in the imported file form part of the ontology defined by the importing file. It is therefore possible for a file to be part of multiple ontologies. It is an ontology in its own right plus it forms part of all the ontologies in which it is included.

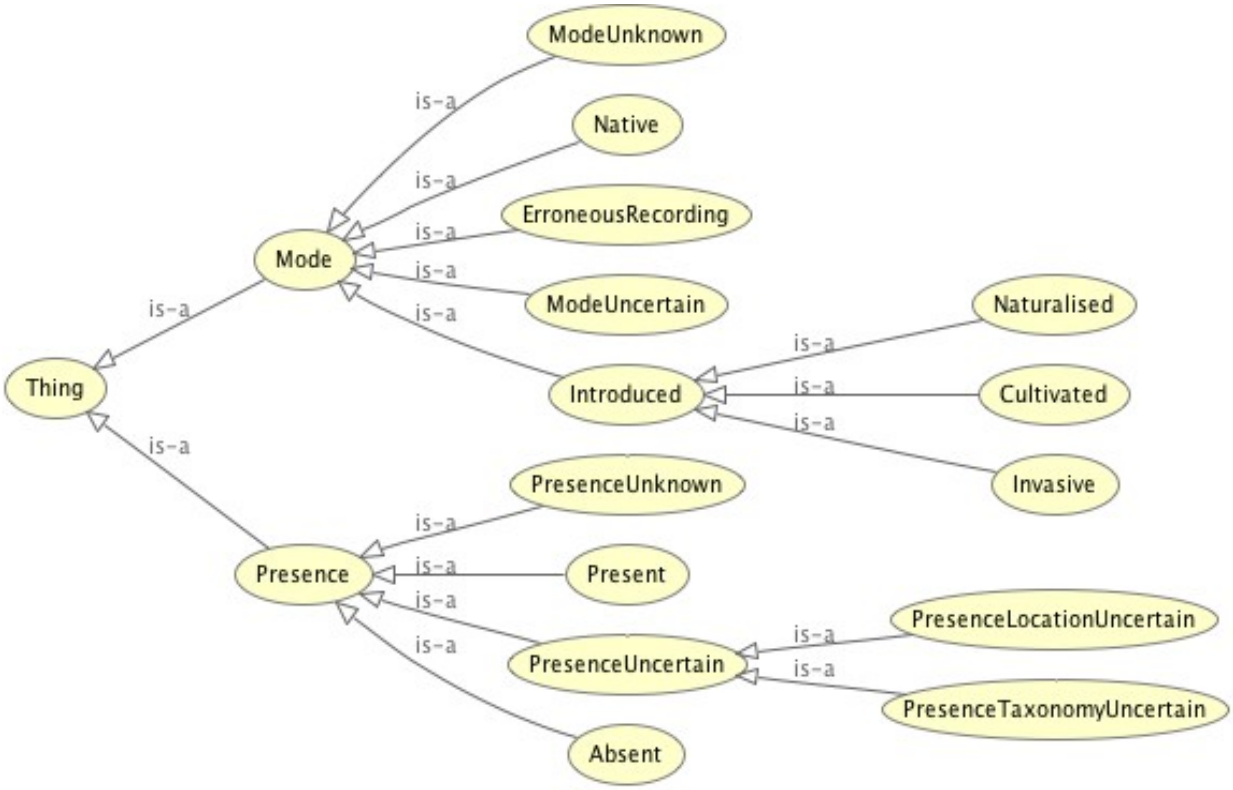
An ontology file was created for each the three databases. An OWL class was declared for each occurrence status. None of the file referenced any of the other files. No semantically significant class hierarchy was created for the statuses presented. This meant they could be included in other

ontologies as required without entailing anything beyond their existence.

As POSS is not used in the three core databases it was not considered during the OWL modelling phase. For the E+M database only the exported codes were considered.

A fifth ontology was created that defined a hierarchy of classes representing the more generic notions believed to be represented in the database statuses, such as presence, absence, native, introduced etc. These generic classes fell into two hierarchies one rooted in the class presence (subclasses of which included present, absent, unknown) and the other rooted in the class mode (with subclasses including native, introduced, cultivated, unknown).

Figure 1: Generic ontology of terms



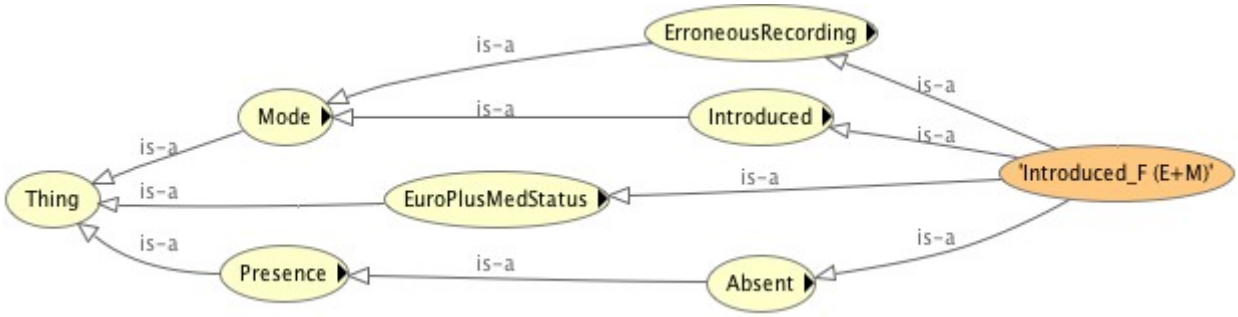
Joining ontologies were then created that joined the database ontologies to the generic terms ontology. These ontologies imported both the relevant database ontology and the generic ontology and asserted OWL equivalent class relationships between the database terms and the generic terms. Doing it this way allowed for the possibility of other ontologies being created that mapped between generic terms and the database terms differently. Finally an ontology including the whole set was created.

Equivalent class are two classes that have the same class extension or membership (i.e., both class extensions contain exactly the same set of individuals). For each of the database ontologies a mapping was defined that consisted of defining an intersection class between the Mode and Presence parts of the generic ontology. As an example. “Introduced F” (interpreted as absent but reported as introduced present in error) from E+M is defined as having the equivalent class that is the intersection of “Absent and Introduced and ErroneousRecording” stated in the Manchester OWL syntax. The same thing expressed as RDF/XML might look like this:

```
<owl:Class rdf:about="em:Introduced_F">
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <rdf:Description rdf:about="generic:Absent"/>
        <rdf:Description rdf:about="generic:ErroneousRecording"/>
        <rdf:Description rdf:about="generic:Introduced"/>
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</owl:Class>
```

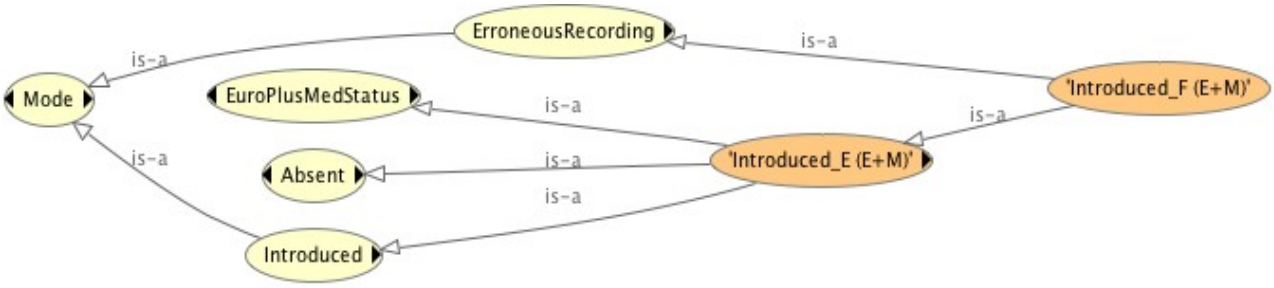
Meaning that any occurrence record that is a member of the generic:Absent occurrence record class **and** the generic:ErroneousRecording class **and** the generic:Introduced class is also a member of the em:Introduced_F and, *visa versa,* a member of the em:Introduced_F class must be a member of all the other three. This is illustrated in figure 1.

Figure 1: Asserted class hierarchy for E+M Introduced F.



Because these assertions are defined in OWL DL it is possible to use a reasoner to elucidate further relationships that were not asserted when constructing the mapping between the two ontologies. The asserted relationships using the Fact++ are shown in Figure 2. This illustrates that, according to this ontology mapping, any Introduced F occurrence is also an Introduced E (Introduced Extinct) which isn't immediately apparent until it is considered that Introduced E is defined as any occurrence record that is “Introduced and absent” i.e. a subset of “Absent and Introduced and ErroneousRecording”. This makes sense when we consider that the only difference between these two notions is that in one the taxon was really present before becoming extinct whilst in the other the taxon was only thought to be present before becoming extinct.

Figure 2 Inferred Relationships for E+M Introduced F.



Running the reasoner against the “all” ontology produced fifty four such “cryptic” subclass relationships involving the twenty database statuses and sixteen classes in the generic ontology. Analysing the combined ontology was very complex and repeatedly raised questions as to whether

the meaning of the original statuses was fully understood.

All the ontologies created are available for download[7].

Conclusions

We started with a four vocabularies. We built a system for reconciling a subset of these vocabularies using OWL DL. This resulted in a series of interrelated ontologies that formed a single complex ontology that appeared to summarise the domain but also appears highly complex. Although the results were understandable on a small scale i.e. looking at parts of the ontologies such as those illustrated in Figures 1 and 2 it was complex and difficult to understand on a larger scale. This would make it awkward to consult a wider audience, who weren't comfortable with OWL DL, as to whether the ontology was correct. Certainly any attempt to take this work further would require commitment of a large amount of resources in the form of modelling experts and domain experts and that commitment must be justified.

Ultimately we are not modelling biochemical pathways here – the result of which might lead to a greater understanding of the biology of an organism. We are modelling a 'mess' that has been created as an artefact of the language used to summarise observational data. When the actual usage of the complex occurrence terms in E+M (Tables 3, 4, 5 & 6) it can be seen that the majority of terms are not used and over 90% of records are covered by just three terms.

These four vocabularies are not alone, there are many other status occurrence vocabularies (e.g. British Ornithologists' Union Species Categories[8]) that could be added. To continue expanding the ontology to account for all of them would increase its complexity and wouldn't ultimately add to our knowledge of biology – possibly only the semantics of human language.

In this case, and possibly in other cases, construction of formal mapping mechanisms between vocabularies would appear to be redundant.

Recommendations for PESI

It is the opinion of the author that it would be better to start with a clean slate rather than try and model existing vocabularies with a complex ontology. A single list of terms should be **defined.** Each term should be as semantically unambiguous as possible – only defining one aspect of an organisms occurrence and not overloading the meaning. As and example “Native” should not be taken to also imply “Present”. If an author wishes to imply that it is present and native both terms should be used.

There should be no form of hierarchy associated with these terms as it implies levels of generality that will not be accepted across different applications and will lead to contradictions. e.g. Placing a “Transit Migrant” as a subclass of “Present” means that all applications have to treat it as present and plot it on distribution maps as such when some applications may only consider “Breeding Migrant” as “Present” and “Transit Migrant” as a subclass of “Absent”. A class hierarchy should therefore not be defined *a priori*. If an author considers an occurrence to fit the description of “Transit Migrant” the record is scored with that term. If it fits the description of “Present” then it is also scored with that term. It is up to the consuming application to decide, in their terms of reference, whether the definition of “Present” warrants “Transit Migrant” to be treated as a subclass of it.

7 http://www.hyam.net/publications/species_occurrence_ontologies.zip
8 http://www.bou.org.uk/reccats.html

The terms should not imply change over time (e.g. “Formerly” or “Extinct”). If it is necessary to describe change through time then a series of dated occurrence records should be created. e.g. “Extinct” could imply loss in modern, historic, prehistoric or geological time scales. All three have very different implications for how the occurrence record might be interpreted.

The terms should be thought of as tags that occurrence records are labelled with. Each record may have one or more, possibly contradictory tags. Each contributing database should use these terms internally or should map their existing terms to these agreed ones manually at export. Table 8 defines a candidate list of terms with example non-overloaded definitions. This list could be extended by adding more terms.

Table 8: Candidate list of Unambiguous Terms (these definitions will need refining)

Term	**Definition**
Absent	There is sufficient evidence to assert that the organism **does not** occur in the region.
Present	There is sufficient evidence to assert that the organism **does** occur in the region.
Native	The organism either evolved in this region or arrived by non-anthropogenic means.
Introduced	The organism arrived in the region via an anthropogenic mechanism or mechanisms.
Naturalised	The organism reproduces naturally and forms part of the local ecology.
Invasive	The organism is having a deleterious impact on another organism, multiple organisms or the ecosystem as a whole.
Managed	The organism maintains its presence through intentional cultivation or husbandry.

Acknowledgements

This work formed part of the Work Package 4 of PESI. PESI is funded by the European Union 7th Framework Programme within the Research Infrastructures programme. Contract no. RI-223806. Period 2008-2011.

I would like to express my gratitude to the current curators of the databases named as well as the numerous expert contributors to these databases. I would also like to thank those who devised the various occurrence vocabularies discussed and acknowledge just how hard the task of creating these vocabularies is.